DOCUMENT RESUME

ED 221 588 TM 820 623

AUTHOR Peters, Eileen
TITLE Supplement to Writing Assessment Manual. Basic Skills Improvement Policy: Implementation Guide 2.
INSTITUTION Massachusetts State Dept. of Education, Boston. Bureau of Research and Assessment.
PUB DATE Jan 82
NOTE 36p.; For related document see ED 203 376.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Standards; Achievement Tests; *Educational Assessment; Guidelines; *Holistic Evaluation; Program Evaluation; *Scoring; State Programs; Student Evaluation; *Writing Evaluation
IDENTIFIERS *Basic Skills Improvement Policy (Massachusetts)

ABSTRACT

This 1982 supplement is a further elaboration to the Writing Assessment which was distributed in 1981 to assist schools in assessment of writing by detailing scoring procedures, standard-setting, and the reporting of results. The supplement deals with retesting, alternative methods for setting standards, establishing score equivalency on writing samples, and maintaining standards. Holistic scoring is an efficient, reliable and direct method of evaluating overall writing ability. Using student writing samples, teachers judge a piece of writing for its total effect relative to other papers. Since absolute scoring methods are not used, the actual quality of the writing each score represents will depend on the range of writing a given group of students produces on a topic. After the minimum district standard is determined, the corrected papers must be re-evaluated in terms of the district's specific writing objectives or definition of a minimally accepted paper. By defining a minimally accepted paper a district can keep track of students' writing performance changes over time. Changing writing topics is discussed, and answers to questions concerning holistic scoring and standard setting are included. Sample rating form, a table for determining sample size, and information regarding using the state writing tests are appended. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED221588

Basic Skills Improvement Policy
Implementation Guide #2

Supplement
to
Writing
Assessment
Manual

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

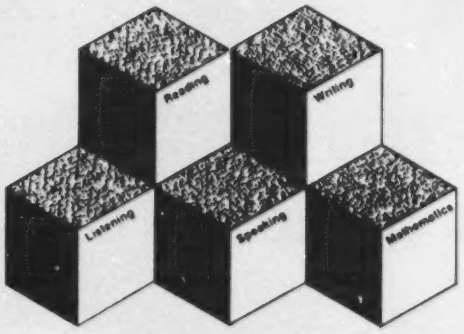
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

C. D. Bills

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."



TN 820 623

MASSACHUSETTS DEPARTMENT OF EDUCATION

JANUARY 1982

MASSACHUSETTS BOARD OF EDUCATION

Donald R. Walker, Chairperson

John W. Bond
Millie Clements
Joseph P. Crimmins
James R. Grande
James L. Green
Howard A. Greis
Charles T. Grigsby
Mary Ann Hardenbergh
Armando Martinez
Edwin Rossman
Dorothea A. Zanetti

John H. Lawson, Commissioner of Education

BUREAU OF RESEARCH AND ASSESSMENT

Allan S. Hartman, Director
Matthew H. Towle
Dorothy F. Earle
Richards S. Williams
Betty Hancock

with the assistance of
Elaine A. Murphy

The Massachusetts Department of Education insures equal employment/educational opportunities/and affirmative action, regardless of race, color, creed, national origin, or sex, in compliance with Title IX, or handicap, in compliance with Section 504.

BASIC SKILLS IMPROVEMENT POLICY

Supplement

to

WRITING ASSESSMENT MANUAL
(Implementation Guide #2)

Bureau of Research and Assessment
Department of Education
31 St. James Avenue
Boston, Massachusetts 02116

January, 1982

Copyright © 1982 Massachusetts Department of Education

This publication is restricted for use in Massachusetts schools for the specific purpose of implementing the Massachusetts Basic Skills Improvement Policy and Regulations and may be duplicated by Massachusetts schools for that purpose. Duplication for use for other purposes requires written permission from the Massachusetts Department of Education.

FOREWORD

This Supplement is an elaboration of two earlier Department of Education publications sent to school systems in 1981: Writing Assessment Manual (Implementation Guide #2) and Standards Setting Manual (Implementation Guide #3). The Supplement focuses on establishing and maintaining equivalent writing standards and addresses other issues raised since the first manuals were written.

Readers of this Supplement will find it helpful to be familiar with both manuals mentioned above and to have each at their disposal. (The Supplement makes reference to specific pages in each earlier publication.) Copies were sent to all School Superintendents and Basic Skills Coordinators, as well as distributed to those attending various training workshops.

Finally, it should be noted that all numerical data are cited for illustrative purposes only and should not influence local performance standards. These standards are to be determined independently as called for in the Regulations for Implementation of the Policy on Basic Skills Improvement.

CONTENTS

		Page
I.	SETTING MINIMAL WRITING STANDARDS: CONTRASTING PAPERS METHOD	1
	How the Contrasting Papers Method Works	1
	How the Method is Carried Out	2
	Selecting the Judges	2
	Preparing for the Rating Session	2
	Conducting the Rating Session	3
	Constructing Cumulative Frequency Distributions	4
	Calculating the Possible Classification Errors	4
	Estimating the Standard	6
	Reminder List for Contrasting Papers Method	7
II.	ESTABLISHING SCORE EQUIVALENCY AND MAINTAINING STANDARDS FROM YEAR TO YEAR	8
	Limitations of Conventional Holistic Scoring	8
	Maintaining Consistent Standards	9
	Conventional Holistic Method: Procedures for Maintaining Standards	10
	An Alternative to Conventional Holistic Scoring	11
	Alternative Holistic Method: Procedures for Maintaining Standards	12
	Setting Standards Using Different Holistic Scoring Procedures	15
III.	CHANGING WRITING TOPICS, TESTING AND RE-TESTING	16
	Testing the Entire Population	17
	Administering a Make-Up Test and Re-testing	18
IV.	ANSWERS TO QUESTIONS ABOUT HOLISTIC SCORING AND STANDARD-SETTING	21
	APPENDICES	24
	Appendix A Sample Rating Form	
	Appendix B Table for Determining Sample Size	
	Appendix C Using the State Writing Tests	

I. SETTING MINIMAL WRITING STANDARDS: CONTRASTING PAPERS METHOD

This chapter of the Supplement describes one set of procedures for setting minimal standards for writing samples that have been scored holistically using conventional holistic scoring methodology. Other standards-setting methods are discussed in the Writing Assessment Manual (Implementation Guide #2), Chapter VI, "Standards-Setting." You are advised to read that chapter as well as this one before finally selecting a method for setting writing standards.

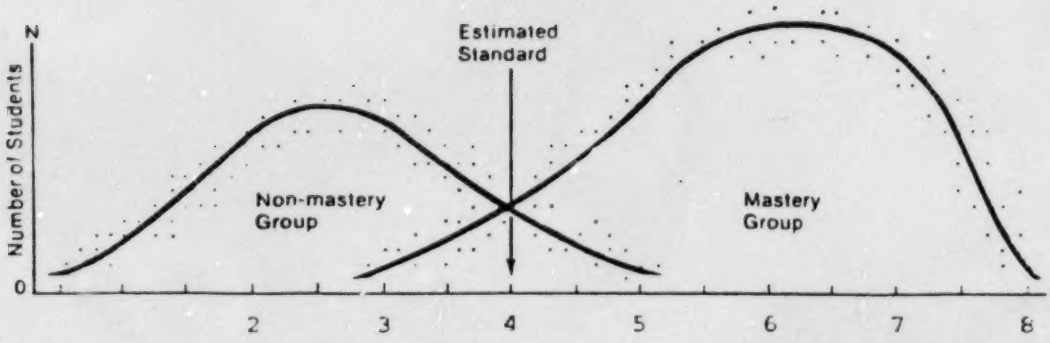
HOW THE CONTRASTING PAPERS METHOD WORKS

First, to obtain consistent results using this method, you need about 100 papers on the same topic. As a rule of thumb, do not drop below 90 papers. Although there is nothing magical about the number 90, it represents a reasonable lower limit for obtaining consistent results. The second requirement is that the method be applied only to holistically scored papers.

In applying this method, a panel of judges rates a sample set of previously scored writing samples as above or below mastery level -- representing at least a minimally competent writing performance or failing to meet that standard. It is important that ratings be independent of scores; therefore, readers should not be included on panels, and judges should not know the scores that papers originally received.

After the sample papers are rated, the final holistic scores obtained by students rated as masters are compared with the final scores achieved by non-masters. As shown in Exhibit 1 (page 2), the standard is generally set at the point where the two groups of scores intersect.

Exhibit 1



HOW THE METHOD IS CARRIED OUT

Selecting the Judges

The judges should be a group of persons who are qualified to decide whether writing samples are or are not minimally acceptable. Teachers of English at the appropriate grade levels are an obvious choice, but you should also consider using employers, school committee members, parents, and community representatives. The more judges, the better their combined judgment will represent the judgment of all persons who could have been judges. You should use at least five judges, and more if you can get them. However, the process becomes cumbersome with too many judges. Twenty would be a reasonable upper limit, for practical reasons. Remember, do not select those who scored the papers.

Preparing For The Rating Session

1. Schedule a work session of about four hours for discussing the writing objectives, reviewing the assignment, and reading and rating approximately 100 papers.

2. Arrange for a comfortable well-lighted meeting room with sufficient table or desk space.

3. Obtain enough copies of each student paper so that a complete set can be distributed to each judge. To ensure anonymity, check to see students' names do not appear on the papers. Also, be sure that scores are not on the papers, and that the papers are scrambled before you photocopy them.

4. Make copies of the writing objectives and the writing assignment for each judge.

5. Make a copy of the recording form (see Appendix A).

Conducting the Rating Session

1. As the first step, have the judges review your basic skills objectives, and make sure they understand the writing assignment. Then, ask them to describe in their own words, the lowest level of performance (in terms of the objectives) they would consider acceptable. Discuss what your district means by a minimally competent writing performance. (You may wish to tape-record this session for future reference.)

2. Have the judges read the first student paper and consider it in terms of your district's writing objectives. Ask the judges, "How many think this paper represents a minimally acceptable writing performance? How many do not?" Then ask one judge who thought the paper was acceptable to explain why, and ask one judge who thought the paper was not acceptable to explain why. Then tell the judges they may change their choices if they want, and ask them to vote again. Record the final vote on the rating form (Appendix A).

3. Table 1, based on hypothetical data, shows all the information you will have for each paper that is scored and then rated. Assume that the thirteen papers included in Table 1 were scored holistically by English teachers and then rated by ten judges, none of whom had scored the papers.

   Column 1 lists all thirteen papers, with each paper identified by code to preserve anonymity. Column 2 presents the final holistic score (the sum of two readers' scores) for each paper listed. Columns 3 and 4 show the number of acceptable and unacceptable ratings each paper received. Paper GG, for example, a paper with a final score of 6 (on a scale of 2 to 8) was rated as acceptable by nine of the ten judges (column 3), and judged unacceptable by one judge (column 4). No one thought the paper was questionable -- not clearly acceptable or unacceptable.

4. To summarize, after the ratings are complete, you will have two pieces of information for each student paper included in the sample -- an holistic score (the relative quality of the paper as determined by two independent readers) and a judgment as to the paper's acceptability (as determined by a panel of judges). You can now combine the scores and ratings and set a minimal standard.

Constructing Cumulative Frequency Distributions

1. First, make a column listing the final holistic scores a paper could receive (see the first column of Table 2 on the facing page). Start with the papers rated "acceptable," and record the number of acceptable papers at each final score point. Referring to the hypothetical data in Table 1, of the seven papers that were judged clearly acceptable, three (M,H, and Q) were scored 8. Therefore, in Table 2, the number 3 is entered next to score 8. (See the second column of Table 2). Continue recording the number of acceptable papers at each score point until you have accounted for all the papers judged acceptable.

   Now, in the column headed "Number below score point" (Table 2) record the number of acceptable papers that received scores lower than the score listed in the first column. In the hypothetical case, for example, three of the seven acceptable papers received the score 8. Therefore, four papers (seven minus three) received scores below 8. (See the third column of Table 2.) Similarly, if two acceptable papers received final scores of 6, the number "below score point" would be two (seven minus three receiving 8, minus two receiving 6).

2. Now, turn to the papers rated "unacceptable" and record the number of unacceptable papers at each final score point. Referring to the hypothetical data in Table 1, of the six papers that were rated unacceptable, one paper (paper MM) was scored 5. Therefore, in Table 2, the number 1 is entered next to score 5 in the column headed "number of papers at score point." Continue recording the number of unacceptable papers at each score point until you have accounted for all the papers judged unacceptable.

   Now, in the column headed "number at or above score point, record the number of unacceptable papers that received either the final score (listed in the first column) or a higher score. In the example (Table 2), three unacceptable papers were scored 4: therefore, the number 4 is recorded next to score 4. The number 4 accounts for the three unacceptable papers receiving 4s, plus the one unacceptable paper scored 5.

Calculating the Possible Classification Errors

For each final score, add the number of acceptable papers receiving lower scores and the number of unacceptable papers receiving higher scores or the same score (see column 6 in Table 2). For example, for score 5, there are two possible misclassification errors (one acceptable paper with a score below 5 plus one unacceptable paper with a score "at or above" score 5.

Table 1. RATINGS OF THIRTEEN WRITING SAMPLES
(Ten Judges)

Paper Code	Final Holistic Score	Clearly Acceptable (N of Ratings)	Clearly Unacceptable (N of Ratings)	Questionable (N of Ratings)
GG	6	9	1	
K	6	8		2
R	3		10	
MM	5		10	
PP	2		9	1
X	5	7		3
P	4	1	7	2
N	4		10	
J	4	10		
M	8	10		
H	8	10		
Q	8	10		
S	4	1	9	

Table 2. POSSIBLE MISCLASSIFICATIONS
(Thirteen Papers)

Final Holistic Score	Papers Rated Acceptable (N=7)		Papers Rated Unacceptable (N=6)		Possible Classification Errors
	Number at Score Point	Number Below Score Point	Number at Score Point	Number at or Above Score Point	
8	3	4	0	0	4
7	0	4	0	0	4
6	2	2	0	0	2
5	1	1	1	1	2
4	1	0	3	4	4
3	0	0	1	5	5
2	0	0	1	6	6

Estimating the Standard

A reasonable estimate for the standard is the score at which possible classification errors are minimized. Thus, for the hypothetical data in Table 2, scores 5 and 6 result in the fewest total misclassifications. Either score might be a reasonable estimate of the minimal standard. Before deciding on a standard, you should read pages 32-40 of Implementation Guide #3.

REMINDER LIST FOR CONTRASTING PAPERS METHOD

- Except for corroborative purposes, do not use this method unless you have at least 90 papers on the same topic.
- Be sure that the judges do not know the scores papers received when they rate them.
- Schedule sufficient time for the rating session.
- Clearly describe the standards-setting procedures.
- Allow enough time to discuss the district's basic skills objectives and enough time to consider the mastery level represented by each paper you rate.
- Allow for discussion of different viewpoints.
- Double check all calculations.

CONTRASTING PAPERS METHOD

Advantages

1. Two independent sets of judgments are collected on students' mastery status.
2. Procedures are similar to those used for setting mathematics and reading standards, and thus may be familiar to various audiences.
3. The method takes questionable or borderline ratings into account.
4. The method can accommodate participation by a broad cross-section of community members as judges (e.g., teachers, administrators, parents, and students).

Disadvantages

1. For consistent results, approximately 100 writing samples, responding to the same topic are needed.
2. A separate panel--one that does not include readers who scored the papers in question--must be named.
3. It may take longer to discuss and rate 100 papers than to reconsider the 2s, 3s, 4s, etc., in order to select a score point representing a minimal performance level.

II. ESTABLISHING SCORE EQUIVALENCY AND MAINTAINING STANDARDS FROM YEAR TO YEAR

This chapter explains why it is necessary to set new writing standards each time the writing assignment is changed or a new group of students is tested, and what can be done to maintain consistent standards from year to year or from one test administration to another. The following questions are addressed:

- Why must new standards be set yearly, even if the writing assignment is not changed?
- How can a district introduce new topics or administer the same topic to different groups of students and still be reasonably confident that the new standard for a minimally competent writing performance is comparable to the one set in previous years?
- How can score improvements be demonstrated?

LIMITATIONS OF CONVENTIONAL HOLISTIC SCORING

In a conventional holistic scoring session, scoring criteria are established by having readers compare sample papers (training papers) with each other. The readers discuss the topic on which students wrote. They then evaluate training papers which have been selected by the Chief Reader as representative of the range of writing produced in the administration of a given topic. To repeat: readers compare papers to each other and not to absolute standards or external criteria.

The scoring criteria and, thus, the scores students obtain are derived from the range of writing produced on a given topic by a particular group of students, as evaluated by a particular group of readers. In other words, three important variables interact to influence score outcomes: the writing task itself, students' writing skills and motivation, and the particular confluence of readers' experiences.

Changes in any of these variables are likely to affect results. Since changes are bound to occur from year to year, there is no reason to believe that a given score (3, for example) will always represent the same level of writing ability. If your minimal standard for the essay was originally set at 3 (on a final score scale of 2-8), you will not know whether this performance standard continues to represent the same quality of writing, unless you evaluate papers analytically. In order to verify that the same scores are consistent from year to year, you need a systematic way of comparing "new" papers with "old" ones.

MAINTAINING CONSISTENT STANDARDS

Consistency for a single round of standard-setting and equivalency from year to year cannot be achieved unless fairly detailed descriptive guidelines are prepared and followed. That is, the characteristics of an unacceptably written paper (in terms of your district's writing objectives) must be stated explicitly and precisely. In order to sharpen contrasts and improve the reliability of results, an acceptable writing performance should be described as well. If, in addition to maintaining comparable minimal standards, you want to compare performance by score category (e.g., determine changes in the percentage of students receiving 2s over a three-year period), you should develop descriptive criteria for major score points (4, 3, 2, and 1).

Conventional Holistic Method: Procedures for Maintaining Standards

Year One

1. The Chief Reader and Assistant Chief Reader select training papers -- papers representing the range of student writing on a particular topic. As they are trained, readers will identify some of these papers as clear 4s, 3s, 2s, and 1s. (Other training papers will not be as easy to categorize and will receive split scores, i.e., 3/2, or even discrepant ones.) Save sample training papers that readers clearly identified as 4s, 3s, 2s, and 1s. You should have three or four papers in each category.

2. After the scoring session is over, develop written descriptions of each point on the score scale (4, 3, 2, and 1), as that score point is exemplified by the sample papers. To do this, read over the sample papers in a given score category, asking yourself: "What characteristics do these papers share? What distinguishes them from sample papers illustrating another score point? How do the 4s differ from the 3s, the 3s from the 2s, etc." A hypothetical definition of a score point might look something like this:

> EXAMPLE
>
> Score 1
>
> Papers reveal little control of written English. Writing is not properly paragraphed, and the sentences are skeletal. The necessary information is not conveyed to the reader. Few capitals or clear punctuation marks occur and when they do, seem randomly placed. The spelling may be worse than one would expect at this grade level.

Year Two

1. Re-read year-one sample papers selected to illustrate the four score points, and refamiliarize yourself with the written description of each score point.

2. In pulling this year's training papers, choose papers that are as similar as possible to those previously selected as sample 4s, 3s, 2s, and 1s. While you should not expect a one-to-one correspondence of strengths and weaknesses for a given score category, you should strive for as much similarity as possible. You should involve a number of people in this selection process. Use these papers to train the readers to score papers holistically.

3. After papers are scored, save all the training papers that readers thought were clear 4s, 3s, 2s, and 1s. Continue to save your first set of sample training papers as well. As time goes on, you will establish a chain, linking scoring standards from year to year. If you have selected and retained benchmark papers for two years, for example, you can link third-year papers to second-year samples or to samples saved from the first year. By the fourth and fifth years, these procedures for maintaining standards can become quite elaborate.

4. If score points are anchored in this way, scores will be roughly comparable from year to year. This will enable you to use the descriptive method of standard-setting (Implementation Guide #2, pp. 66-67) and still compare results from year to year.

AN ALTERNATIVE TO CONVENTIONAL HOLISTIC SCORING

There is another way to maintain consistent standards from year to year, but this method assumes a variation in holistic scoring procedures. Instead of letting score 1 (scale 4 to 1) represent writing that is relatively weak compared to other writing on the same topic, you reserve score 1 for writing that is not only relatively weak, but unacceptable in terms of your district's minimal writing standards.

Before reading the next section of this Supplement covering procedures for maintaining standards, you should review the alternative holistic scoring method, Implementation Guide #2, pp. 60-63.

Alternative Holistic Method: Procedures for Maintaining Standards

Year One

1. The Chief Reader and Assistant Chief Reader select training papers including at least twenty "lower-half" papers. Some will be very poorly written, but more should show a mix of major weaknesses with some redeeming qualities. During a meeting held well in advance of the actual scoring session, readers identify papers they think are minimally acceptable and those which they regard as unacceptable. These papers, along with a rationale for a proposed standard, are then reviewed by a committee of community representatives. The committee's definition of minimally acceptable and unacceptable writing, along with samples of student writing, guide the selection of below standard training papers. These papers and others like them will serve as examples of papers that readers will be trained to score 1 -- unacceptable.

   While a committee's definition of an unacceptable writing performance (score 1) might be as brief as the example shown on page 10, it could also be more inclusive. The example of empirically derived archetypes presented on page 13 takes various patterns of deficiencies into account, alerting readers to the fact that 1s may not exhibit weaknesses uniformly.

2. After the scoring session is over, save sample training papers that were clearly identified (by readers and the public panel) as unacceptable -- papers that most people agreed were good examples of unacceptable writing. Be sure to keep papers scored 1 that represent a range of unacceptable writing performances. (If you save only the most poorly written papers, you will be artificially lowering your standard for the following year.) You should also keep a few "lower-half" papers that readers and the public panel thought were weak, but minimally acceptable.

EXAMPLE[1]

CHARACTERISTIC 1s

Many have been so short (a few sentences/less than a hundred words) that they have been classified as 1s because the writer simply has given no space for development or organization. What exists is a collection of sentences, not a composition.

Some have been lists. Whether single sentence paragraphs, one long paragraph, or a list format, these 'papers' are still lists and not compositions. Other papers of this type might be referred to as outlines.

Some have had such serious problems in mechanics and/or handwriting that they cannot be read; they must be deciphered.

Some have had a density of errors in spelling and mechanics combined with little development or elaboration.

Some have had such serious problems in organization, particularly logical connections between ideas, that they cannot be read and understood.

Some have had serious problems in syntax and grammar, e.g., persistent lack of inflectional endings for verbs, persistent problems with subject-verb agreement, ungrammatical sentences. (What is meant here is not a difficulty with punctuation, but deep syntactical problems that suggest a lack of understanding of what an English sentence is.)

Some have had such impoverished vocabularies (no words with more than five letters) and such short and repetitive sentences (4-8 words) that the writer's options are seriously limited.

Some take the form of transcribed speech. Spelling, use of colloquialisms, run-on sentences, the use of pauses, and the associative ellipses characteristic of speech are all signs of this kind of transcription.

[1]These archetypical descriptions were developed by Robert Zeeb, Coordinator of English; Newton Public Schools, Newton, Massachustts

Year Two

1. Review your training papers from the previous year along with other materials pertaining to your district's minimal standard.

2. Select training papers that match, as closely as possible, papers that your community previously identified as unacceptable and minimally acceptable. If you changed the topic (and even if you did not), you will not achieve complete comparability between the two sets of papers. Just try to find as many new papers as you can that reflect combinations of characteristics that your district decided represented an unacceptable writing performance. You should involve a number of people in this process. Use these papers (and sample "upper-half" papers) to train readers.

3. After papers are scored, save a range of sample 1s and a few new papers judged weak but minimally acceptable. Save your old training papers -- those reviewed by the community panel -- as well. The following year, repeat the process of using benchmark papers (the original set or the ones identified during the second round) to guide the selection of new training papers.

4. Your minimal standard will be maintained as long as unacceptable (score 1) papers continue to exhibit the characteristics originally associated with a below-standard writing performance as defined by the community panel. Although it is not necessary for the public to review this standard yearly, you should provide for periodic consideration of minimal performance standards.

SETTING STANDARDS USING DIFFERENT HOLISTIC SCORING PROCEDURES

CONVENTIONAL HOLISTIC

Advantages	Disadvantages
1. Method is more familiar than alternative ones.	1. Method is normative and does not, in and of itself, identify writing that is above and below a particular standard.
2. Training and scoring procedures are relatively easy to implement.	2. It is time consuming to score papers and then re-examine them analytically to set standards.

ALTERNATIVE HOLISTIC

Advantages	Disadvantages
1. Standards are set efficiently as papers are scored.	1. Method is less familiar to readers, and it may be more difficult for the Chief Reader to conduct the scoring session.
2. Results are probably more consistent than those obtained by scoring papers in the conventional holistic method and then reviewing them analytically to set standards.	2. Method requires that scorers calibrate to a descriptive standard which may take more training time than conventional holistic scoring.

III. CHANGING WRITING TOPICS, TESTING AND RE-TESTING

Almost everyone would agree that students' scores on a writing sample might be affected by such things as distractions in the classroom at the time of testing, the quality of prior instruction, the experience and characteristics of those judging the writing performance, the weather, and a multitude of other factors. In fact, anything that influences the way students write, how readers score, or a combination of the two, will affect performance ratings. It should come as no surprise then, that the topic or writing assignment students are given will also affect outcomes. Two generalizations follow from this observation:

- The same student may perform differently on one topic (for example, an assignment calling for argument) than on another (for example, one asking for description).
- Similarly, because we cannot assume that topics affect performance in similar ways unless they have been shown to do so, we generally should not treat comparable scores on <u>different</u> topics as if they represented the same performance level. This year's 3 is not necessarily equivalent to last year's 3 if different topics were administered.[1]

The process of anchoring minimal standards, and of using sample papers to define score points and guide the selection of training papers for another round of scoring, will help you to maintain roughly the same basic standards from year to year. Using the same topic will help, too, simply because doing so will make it easier to match sets of training papers.

Despite the convenience of re-administering the same assignments from year to year, there are reasons for changing topics. The first point to consider is whether you are testing the entire population at a given grade level or administering a "make-up" or re-test.

---

[1]Conventional holistic scoring compounds this problem by ruling out absolute performance standards in favor of relative ones: "Compared with other papers I've read, this paper is a 3."

TESTING THE ENTIRE POPULATION

There are two basic reasons for using the same topic(s) for more than one year. It will be easier to link new results with previous performance standards if the topic remains constant, and you must link results in order to compare performance from one year to the next. Furthermore, once you have found topics that work well for you, you may want to repeat them.

But the reasons for switching topics may also be compelling. The importance of test security should be reviewed thoroughly before you decide to repeat topics. Will teachers begin to teach to the assignment? (If some teachers do and others do not, administration conditions will not be uniform, and some youngsters will be at a disadvantage.) Will some students prepare responses in advance, or will they memorize responses developed by others? Even if teachers or students do not behave differently, will the credibility of the administration be questioned? Finally, repetition causes boredom. While student populations will change, the same readers will probably score papers for more than one year. Readers' fatigue and boredom affect scores.

To summarize, there are a number of things to consider before deciding whether to reuse a writing topic or introduce another. If test security is not an overriding concern, it is probably more convenient to repeat topics at least once. Ideally, however, you should begin to field test writing assignments, constructing a pool of topics from which to draw. This practice will minimize the test security problem, reduce boredom, and, most importantly, yield good writing assignments.[1] And, though it will not be as easy to compare performance from year to year, if you change topics, the development of precise score descriptors will help a great deal.

---

[1]The importance of reviewing and field testing writing assignments is discussed in detail in the Writing Assessment Manual (Implementation Guide #2), pp. 17-20.

ADMINISTERING A MAKE-UP TEST AND RE-TESTING

Unless you are especially concerned about test security, use the same topic for a make-up administration to absentees or transfer students. For one thing, if you do change assignments and score papers holistically, you will not be able to compare papers written by absentees or transfers with writing samples produced by other students. And, unless you have a very large number of make-ups, the papers will have to be scored analytically and a cut-off score established for this method. This will mean that the overall testing/scoring conditions are substantially different for the main groups of students and for the absentees or transfers.

Re-testing is more complicated. Three groups of students should be re-tested at the elementary level and must be re-tested at the secondary level.

- Students whose original papers could not be scored;
- Students who previously did not meet your district's minimal writing standard;
- Students for whom you need an additional piece of information, as called for by your district's decision rules. (You might have a rule, for example, that requires an additional writing sample in the event that a minimal competency test performance is at variance with general classwork.)

Students whose papers could not be scored, those receiving a "0" score, could be asked to write on the same topic again or given a different topic. If the same topic is given, the re-test papers can be scored holistically by using the original training papers to establish scoring criteria.[1]

[1]See page 20 of this supplement for a further discussion of this point.

If you change topics and do not have explicit scoring criteria, you must score the re-test papers analytically. Of course, there may be little point in asking a student to address the same topic if he or she could not work with it in the first place. If you suspect that the specific assignment is the problem, change it and score the re-test papers analytically.

Re-test papers should not be scored holistically using conventional holistic scoring methodology (even if there are a substantial number) because we cannot make assumptions about a normal score distribution. We cannot assume, for example, that most of the papers that would be compared with others in this particular batch would represent an "average" writing performance.

If, on the other hand, you are using alternative holistic scoring procedures and have developed precise and complete descriptions of a below-standard writing performance (score 1), you could score these papers holistically.

Students who previously failed to meet your district's standards can be re-tested, most conveniently, at the same time the new group of students is tested. The same assignment would be given, for example, to those who are in the eighth grade this year and to last year's eighth graders who did not meet minimal standards. If you elect this alternative, treat all students' papers in the same way. Do not, for example, distinguish between the bulk of the writing samples and the re-test papers when you train readers or ask them to score papers. Readers should not know which papers they are scoring.

If, for some reason, you decide that students who must be re-tested should write on the same topic they were originally assigned, even though you are not giving this assignment to all students again this year, the re-test

papers must be scored separately. In this event, you should not score these papers holistically, unless you co rect for the score distribution problem noted above. You can only do this if you have a complete set of last year's papers, including those papers that were used to train readers.[1] You will have to decide whether you have enough re-test papers to merit the additional clerical and re-scoring work. Alternatively, you can score these papers analytically if you have well-developed scoring rubrics for this process.

A final summary point on re-testing. Since the topic you select and the precision of your score descriptions (standards) will affect scoring possibilities, the entire testing and scoring process should be thought through before writing assignments are made.

[1]First cover or erase scores or other evidence of writing quality. Then randomly select a sample of these papers. (See Appendix B for a discussion of sample size.) Mix these papers in with the re-test papers, checking to be sure that papers appear identical so that no reader will be able to distinguish between an original paper and the re-test paper. Use last year's training papers to establish scoring criteria. Score all the papers. And finally, break the code to identify the re-test papers.

IV. ANSWERS TO QUESTIONS ABOUT HOLISTIC SCORING AND STANDARD-SETTING

1. Q. Is it necessary to set new standards every year -- each time papers are scored holistically?

   A. That depends on the scoring procedures you adopt. If you use the alternative holistic method (Implementation Guide #2, pp. 60-63), the standard is set before papers are scored. A "1" paper (final score 2), as described in terms of your district's writing objectives, will be below standard by definition. The rating a paper receives will depend upon whether or not it satisfies these criteria. The rating (above or below standard) will have very little to do with the quality of other papers. Because the standard is built into the scoring process, it can be maintained from year to year. A paper with a final score of 2 (two raters, each assigning a score of 1) will always be below standard. The meaning of this score will not change.

   If, on the other hand, you use conventional holistic scoring procedures, the judgment of a paper's quality or the score it receives does not depend upon a built-in standard of acceptability. When a reader scores a paper, he or she is not saying: "Does this paper meet these criteria of acceptability (the district's writing standard)?" Rather, the reader is saying: "How does this paper compare with other papers I've read in this batch?" Since the meaning of a score will change from year to year (depending on the readers' standards, students' abilities and motivation, and attributes of the topic), the standard-setting process must be repeated every year.

   To summarize, if you use the alternative holistic approach, your standard is set, although from time to time you will want to review your descriptions of a below-standard writing performance. If you use conventional holistic scoring methodology, you cannot assume that last year's numerical standard will be the same as this years: last year's 2, for example, may be equivalent to this year's 3.

2. Q. Is a "0" paper below standard?

   A. No. A "0" means that, for one reason or another, the paper could not be scored. A paper, for example, might be blank. Students with scores of zero will have to be re-tested.

3. Q. If a student wanders off the topic or otherwise fails to fulfill the requirements of the assignment, should the paper be given a "0"?

   A. No. The score "0" is reserved for papers that are so far off the subject that they cannot be compared with other papers. The "0" score says nothing about a paper's quality or the extent to which a student met your district's writing objectives. It says simply that the paper cannot be scored, and the student must be re-tested.

4. Q. Should community representatives be involved in setting your district's minimal standards for writing?

   A. Yes. The reasons for community involvement are the same for writing as they are for reading and mathematics. See Implementation Guide #2, Chapter VI, for a discussion of this issue.

5. Q. Should re-test papers be scored holistically?

   A. The answer to this question depends upon how you structure the scoring process. You should not score a batch of re-test papers holistically using conventional scoring procedures, because you cannot assume they represent a range of writing ability. Rather, you should score these papers analytically or adjust scoring procedures as suggested on pages 19 and 20 of this Supplement.

6. Q. Should the writing test be timed?

   A. The only reason to time the writing test is to insure that all students are writing under the same conditions. If you impose a time limit, you should see that it is adhered to in all classrooms. If you do not set a time limit, you should be sure that all students have sufficient time to finish their work.

7. Q. Should the topic be discussed before students write?

   A. This question is similar to the one on timing. If no one engages in prewriting activities or discusses the topic in advance, all students will be writing under the same conditions. If preparation takes place, it may be difficult to insure standardized testing conditions.

8. Q. Doesn't it make more sense to score papers written by youngsters in the early and middle elementary grades analytically rather than holistically?

   A. Not necessarily. The scoring procedure should relate to the nature of the writing assignment rather than to the age of the students. If you ask students to write a simple sentence and you want to check for the completeness of a thought and the appropriate end punctuation, you would probably want to score the responses analytically. If, on the other hand, you asked students -- even second or third graders -- to organize and express their thoughts on some subject, there would be no reason not to score these responses holistically.

9. Q. Is there more than one way to sc e the letter?

   A. Yes, there are at least three basic alternatives. First, you can score the letter once holistically, considering form and content together. Second, you can score the letter twice, scoring holistically once for form and once for content -- two separate scores. (If you do this, you will need two sets of training papers -- a set for each of the two attributes, and you will need to train readers twice.) Finally, you can score holistically for content and analytically for form. Again, you will have two scores and will have to decide how to combine them, whether to weight them, etc.

10. Q. If we decide to administer the State writing test, is it necessary to draw the letter and essay assignments from the same test form?

    A. No; however, you must include a letter and an essay if you say you are administering the State writing test.

11. Q. Where can I find a complete discussion of reporting requirements?

    A. Reporting requirements, including examples of appropriate and inappropriate ways of combining scores, are included in Appendix C to this Supplement.

APPENDIX A

SAMPLE RATING FORM

RATINGS OF WRITING SAMPLES

(Number of Judges _____)

Paper Code	Final Holistic Score	Acceptability of Writing in Terms of Objectives		
		Clearly Acceptable (Number of Ratings)	Clearly Unacceptable (Number of Ratings)	Questionable (Number of Ratings)

APPENDIX B

TABLE FOR DETERMINING SAMPLE SIZE
FROM A GIVEN POPULATION[1]

N	S	N	S	N	S
5	5	220	140	1200	291
10	10	230	144	1300	297
15	14	240	148	1400	302
20	19	250	152	1500	306
25	24	260	155	1600	310
30	28	270	159	1700	313
35	32	280	162	1800	317
40	36	290	165	1900	320
45	40	300	169	2000	322
50	44	320	175	2200	327
55	48	340	181	2400	331
60	52	360	186	2600	335
65	56	380	191	2800	338
70	59	384	192	3000	341
75	63	400	196	3500	346
80	66	420	201	4000	351
85	70	440	205	4500	354
90	73	460	210	5000	357
95	76	480	214	6000	361
100	80	500	217	7000	364
110	86	550	226	8000	367
120	92	600	234	9000	368
130	97	650	242	10000	370
140	103	700	248	15000	375
150	108	750	254	20000	377
160	113	800	260	30000	379
170	118	850	265	40000	380
180	123	900	269	50000	381
190	127	950	274	75000	382
200	132	1000	278	100000	382
210	136	1100	285	1000000	384

NOTE: N is population size
S is sample size

[1]Robert V. Krejcie and Daryle W. Margan, Determining sample size for research activities. Educational and Psychological Measurement, Vol. 30, 1970, 607-610.

APPENDIX C

USING THE STATE WRITING TESTS

USING STATE TEST FORMS

Local school districts using the state secondary level writing test have five forms of this test available to them: Forms 2, 3, 4, 5, and 6. Each form contains a letter and an essay. Two questions have been raised about the use of state test forms.

> ONE, MUST EACH FORM BE USED AS IT IS OR, FOR EXAMPLE, CAN THE LETTER FROM FORM 2 BE USED WITH THE ESSAY FROM FORM 3?

Schools may mix state test forms as long as each administration of the test contains both an essay and a letter. However, giving two of the same type of writing sample (i.e., two essays) or administering only one sample (i.e., only the letter) invalidates the state test since this would neither fairly nor comprehensively assess the writing skills enumerated in the Regulations for Implementation of the Policy on Basic Skills Improvement.

If a school district wants to use only one topic, it should amend its Secondary Level Basic Skills Improvement Program Plan to indicate that it plans to use a locally developed writing test and will abide by the requirements noted for writing tests.

> TWO, MUST ALL STUDENTS BE ADMINISTERED THE SAME FORM OF THE WRITING TEST DURING THE INITIAL TESTING OR MAY SOME BE ADMINISTERED ONE FORM AND OTHERS ANOTHER FORM?

All students should be administered the same form of the writing test during initial testing. The state writing tests are designed to be scored either holistically or analytically. The reliability of this type of scoring is based on the use of identical writing topics with a group of trained raters. Scoring is more complicated and takes longer if different topics are used. Additionally, this will create major complications in setting cut-off scores.

ADMINISTERING STATE WRITING TESTS

The Test Administrator's Manual accompanying the writing tests assumes two things: students will be able to complete both writing exercises in 70 minutes and both exercises will be administered in one sitting. Three questions have been raised about these assumptions.

ONE, MAY STUDENTS BE GIVEN MORE THAN 70 MINUTES TO COMPLETE BOTH WRITING EXERCISES?

Students may be given more than 70 minutes as long as the school district applies this uniformly to all students taking the state writing test. The state writing test is a power test; the goal is not to assess speed in writing. The Department's experience, however, is that almost all students will have provided a sufficient sample of their writing during 70 minutes for purposes of holistically or analytically scoring these samples.

TWO, MUST THE WRITING TEST (BOTH LETTER AND ESSAY) BE GIVEN IN ONE SITTING?

Although the state writing test has been designed for administration in one sitting, schools may, if they wish, given one part of the test (i.e., letter) in one test session and the second part (i.e., essay) in another test session. To do this schools should provide students with only that part of the test to be taken at that particular testing session.

THREE, MAY STUDENTS PREPARE AN INITIAL DRAFT OF THEIR WRITING IN ONE TEST SESSION AND THEN RETURN TO PREPARE A FINAL COPY TO BE SUBMITTED AT A SECOND TEST SESSION?

Students may prepare initial and final drafts in either one or two sessions as long as a school district standardizes this procedure for all students. Students, for example, may be instructed to write an initial draft of the essay in one test session, leave the materials with the test monitor when complete, and then return to revise and prepare final copy in a second test session. Students should not be allowed to take test materials home or return with materials for the second test session as this could invalidate the purpose of such an arrangement. Under this arrangement only the second student draft would be scored.

SETTING AND REPORTING STANDARDS ON THE STATE WRITING TESTS

Currently, many districts are using two Department guides to assist them in setting standards on holistically scored writing samples (Writing Assessment Manual and Standards-Setting Manual). Two questions on setting and reporting standards on the state writing test have been raised.

ONE, HOW DOES A DISTRICT SET A MINIMUM STANDARD FOR WRITING WHEN THE STATE TEST HAS TWO DIFFERENT PARTS?

Any defensible method of setting a standard for writing has to start with setting a cut-off score on each writing topic. This is logical, because any defensible approach includes looking at samples of student writing in relationship to the objectives to be measured.

After setting a cut-off score on each topic, a district has four choices:

1. The two scores can be added together to form a combined score, and students required to reach that combined score to be considered above standard. There are two problems with this approach. First, students might do very well on one topic and very poorly on the other. Second, if a student does not meet minimum standards and has to be retested at a later date, the student would have to take both writing topics again and both would have to be scored, creating more administrative work for the district. Also, by looking at a combined score rather than individual scores for the different topics, the district loses information that could be of instructional value.

2. The two scores can be averaged to form one required cut-off score and students required to reach that average score to be considered above standard. This approach has the same limitations stated above.

3. The two scores can be kept separate, and students required to reach the minimum score on the letter or the essay. The main problem with this is that the student could perform well on some of the state writing objectives and very poorly on others, and still be considered above standard. The student's achievement on some of the objectives would not be fully assessed.

4. The two scores can be kept separate, and students required to reach a minimum score on both the letter and the essay. This method has three advantages. First, it assures a thorough coverage of the writing objectives. Second, if a student performs above standard on one topic and below standard on the other topic, the follow-up instruction can be more targeted -- the district need focus only on a limited range of objectives. Third, there is also a major administrative advantage in this approach. If a student reaches the cut-off score on one topic, he/she will have to be reassessed only on the other topic, rather than on both topics. This will save testing and scoring time. Because this method is the most defensible in terms of curriculum and the most efficient use of district resources, it is recommended by the Department.

TWO, HOW SHOULD A SCHOOL DISTRICT REPORT ITS LOCAL WRITING STANDARD ON THE ANNUAL REPORT FORM?

In Section 1 of the early elementary, later elementary, and secondary forms of the 1981/82 Annual Report on Basic Skills Improvement Programs, schools should describe their standard in one of four ways (secondary level examples are used to illustrate):

1. Students required to receive a minimum combined score of 7 out of 16 on the letter and essay to meet local district standards.

2. Students required to receive a minimum average score of 3.5 on the letter and essay to meet local district standards.

3. Students required to receive a minimum score of 4 out of 8 on the letter or 3 out of 8 on the essay to meet local district standards.

4. Students required to receive a minimum score of 4 out of 8 on the letter and 3 out of 8 on the essay to meet local district standards.*

*The numbers are for illustration only. This is an example of the approach recommended by the Department.

While the above example pertains mainly to the state writing test, districts should make appropriate modifications when reporting on locally developed writing instruments or when they supplement the use of writing samples with commercial, multiple-choice writing tests. (See page 3 of Instructions for Completing 1981-82 Annual Report on Basic Skills Improvement Programs.)

MAINTAINING ANONYMITY ON STATE WRITING TESTS

In scoring student writing samples or reviewing samples to set minimum standards, it is important that readers not know the identity of students. Anonymity is essential to ensure reader reliability. To ensure anonymity various coding procedures may be used for identifying papers when they are submitted and/or after they are scored (i.e., giving each student an ID number). One question concerning anonymity on the letter in the state writing test has been raised.

WON'T THE FACT THAT STUDENTS USE THEIR OWN NAMES AND ADDRESSES IN THE STATE WRITING TEST IDENTIFY STUDENTS TO READERS?

While the use of actual names was recommended to the Department by the Writing Committee that advised on the development of the writing tests (to encourage students to take the assignment seriously), there are practical problems with this in preserving anonymity. Three possible ways of maintaining student anonymity are suggested:

1. Eliminate names and addresses on the letter either by instructing students not to include a heading and signature or by using a scissors or magic marker to delete the heading and closing.

2. Instruct the students to use a given name and address which is written on the blackboard. (Use a unisex name such as Chris Peters.)

3. Instruct students to use fictitious names and addresses for themselves.

BASIC SKILLS IMPROVEMENT POLICY

Department Resources - Writing

General

Basic Skills Improvement Policy and Regulations	January, 1979
Measure for Measure - Evaluating Student Writing	December, 1980
Topic Ideas for Assessment Writing at the Elementary Level	January, 1981

Implementation Guides

Assessing Basic Skills Achievement - Guide #1	May, 1980
Writing Assessment Manual - Guide #2	March, 1981
Standards Setting Manual - Guide #3	March, 1981

Assessment Instruments and Information

Massachusetts Test of Basic Skills (Forms 2-6)	1980 & 1981
Supplemental Information on Secondary Level State Writing Test (memorandum)	January, 1981
Massachusetts Assessment of Basic Skills - 1979/80 (Mathematics, Reading, Writing)	September, 1979
1980/81 Statewide Summary of Student Achievement of Minimum Standards in the Basic Skills of Reading, Writing and Mathematics	January, 1982

All the above have been distributed to Massachusetts school systems. Readers should check with the appropriate regional center to see if additional copies of the above are available for examination or distribution.